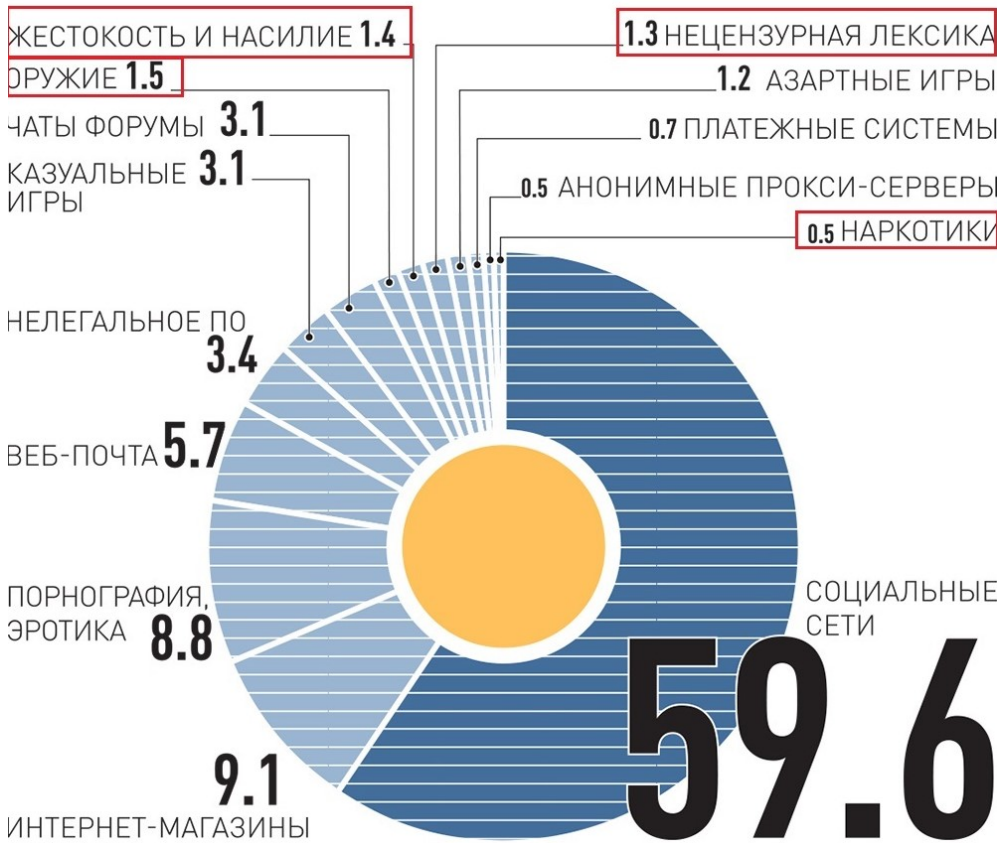
Классификация текстового контента

Александр Смирнов и Федор Жилкин

24.05.2019г

Введение


ЖЕСТОКОСТЬ И НАСИЛИЕ 1.4
ОРУЖИЕ 1.5
ЧАТЫ ФОРУМЫ 3.1
КАЗУАЛЬНЫЕ ИГРЫ 3.1
НЕЛЕГАЛЬНОЕ ПО 3.4
ВЕБ-ПОЧТА 5.7
ПОРНОГРАФИЯ, ЭРОТИКА 8.8
9.1 ИНТЕРНЕТ-МАГАЗИНЫ
1.3 НЕЦЕНЗУРНАЯ ЛЕКСИКА
1.2 АЗАРТНЫЕ ИГРЫ
0.7 ПЛАТЕЖНЫЕ СИСТЕМЫ
0.5 АНОНИМНЫЕ ПРОКСИ-СЕРВЕРЫ
0.5 НАРКОТИКИ
СОЦИАЛЬНЫЕ СЕТИ
59.6

Цели

- Ограничить детей от взрослого текстового контента в интернете
- Получить опыт
  - Бинарная классификация текста
  - Сбор данных для обучения
  - Написание Python-библиотеки
  - Написание расширения для Chrome
  - Написание Python-сервера для приёма запросов

Сравнение с аналогами

- Ограничения на поиск
  - Семейный поиск Яндекс
  - Безопасный поиск Google
- Контентная фильтрация
  - Traffic Inspector
  - Интернет Цензор

Сравнение с аналогами (2)

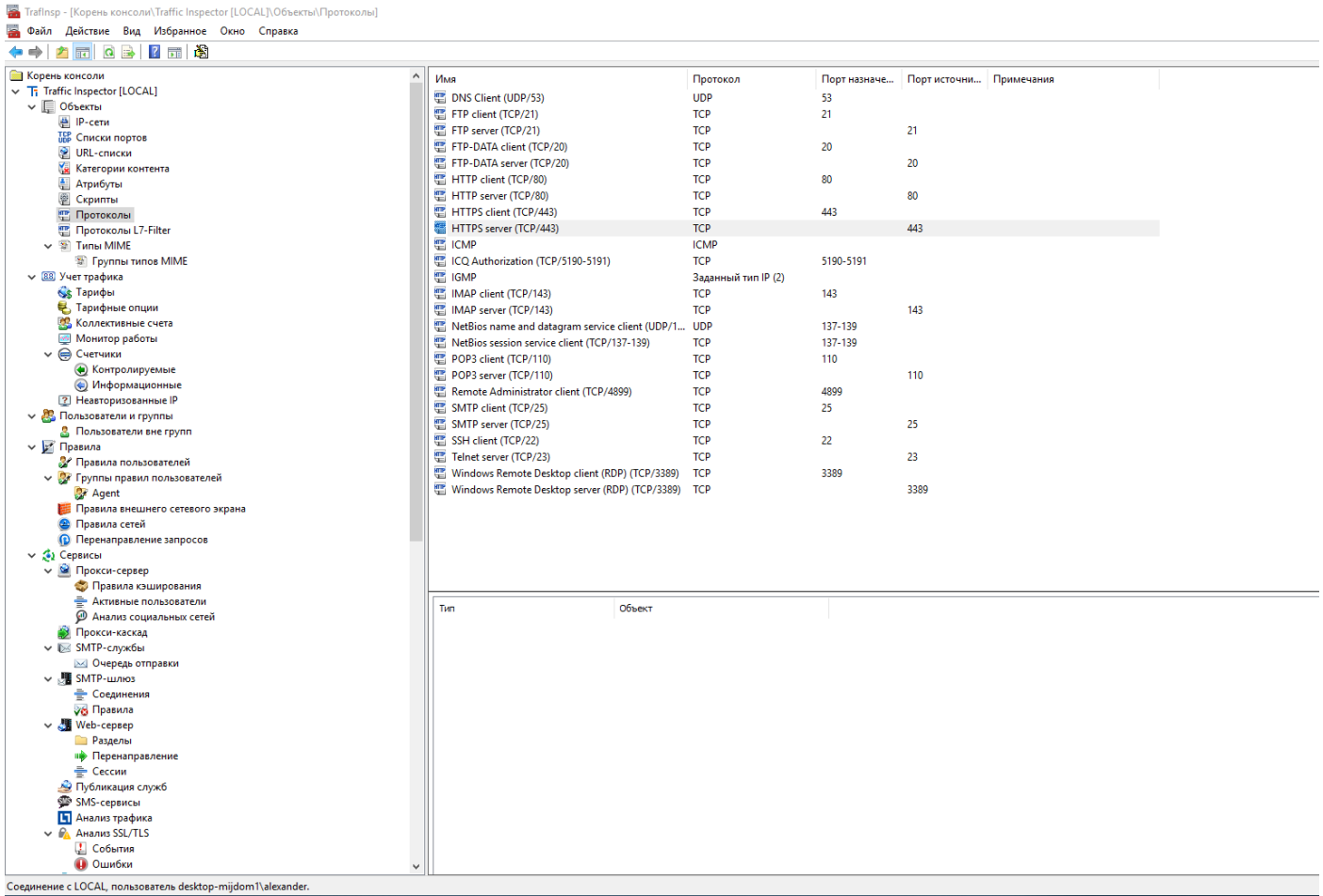


Рис.: Пример интерфейса схожей программы

Задачи

- Провести анализ возможных решений для классификации текста
- Собрать рассказы для взрослых и обычные рассказы
- Написать Python-сервер, использующий обученную модель для ответа на запросы от расширения
- Сделать расширение для Chrome, обращающееся к серверу

Сбор данных

- Рассказы для взрослых берём с сайта ideer.ru
- Рассказы для широкого круга читалей берём с множества сайтов по разным тематикам

Анализ подходов к классификации текста

- Rule-based
- Machine Learning based
- Hybrid systems

Характеристики сравнения эффективности

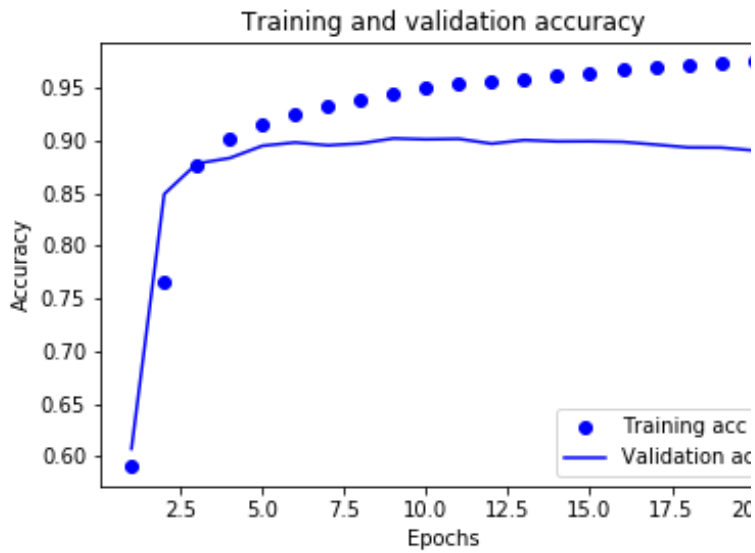
- Accuracy – общая точность классификатора
- Recall – отношение заблокированных взрослых сайтов к общему количеству взрослых сайтов (% классифицированных взрослых сайтов)
- Precision – отношение заблокированных взрослых сайтов к числу всех заблокиронных сайтов (точность блокировки)
- F1 Score – среднее гармоническое между Precision и Recall, для учёта и того, и другого в одной величине

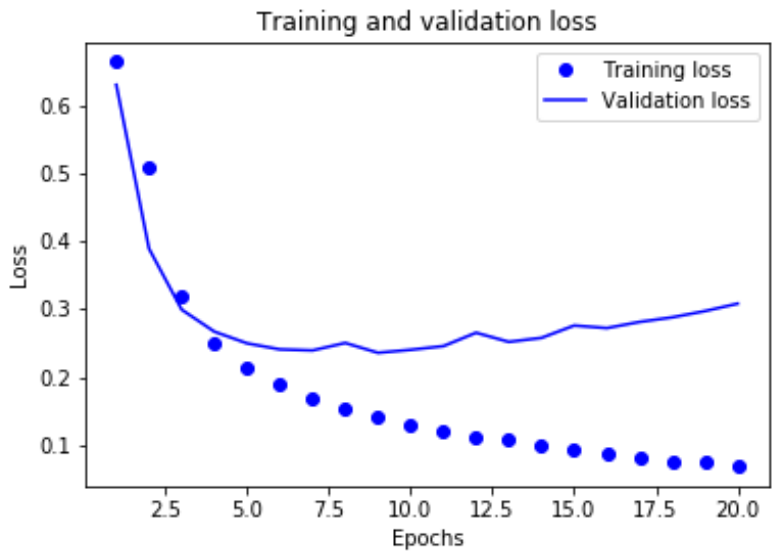
Сравнение

- Random model – случайный выбор блокировать/не блокировать
- Rule-based model – блокировка по списку непотребных слов
- Classifier – 3-х слойная обычная сеть
- Upgraded Classifier – Classifier, из словаря которой были исключены самые частые слова и добавлена ненормативная лексика

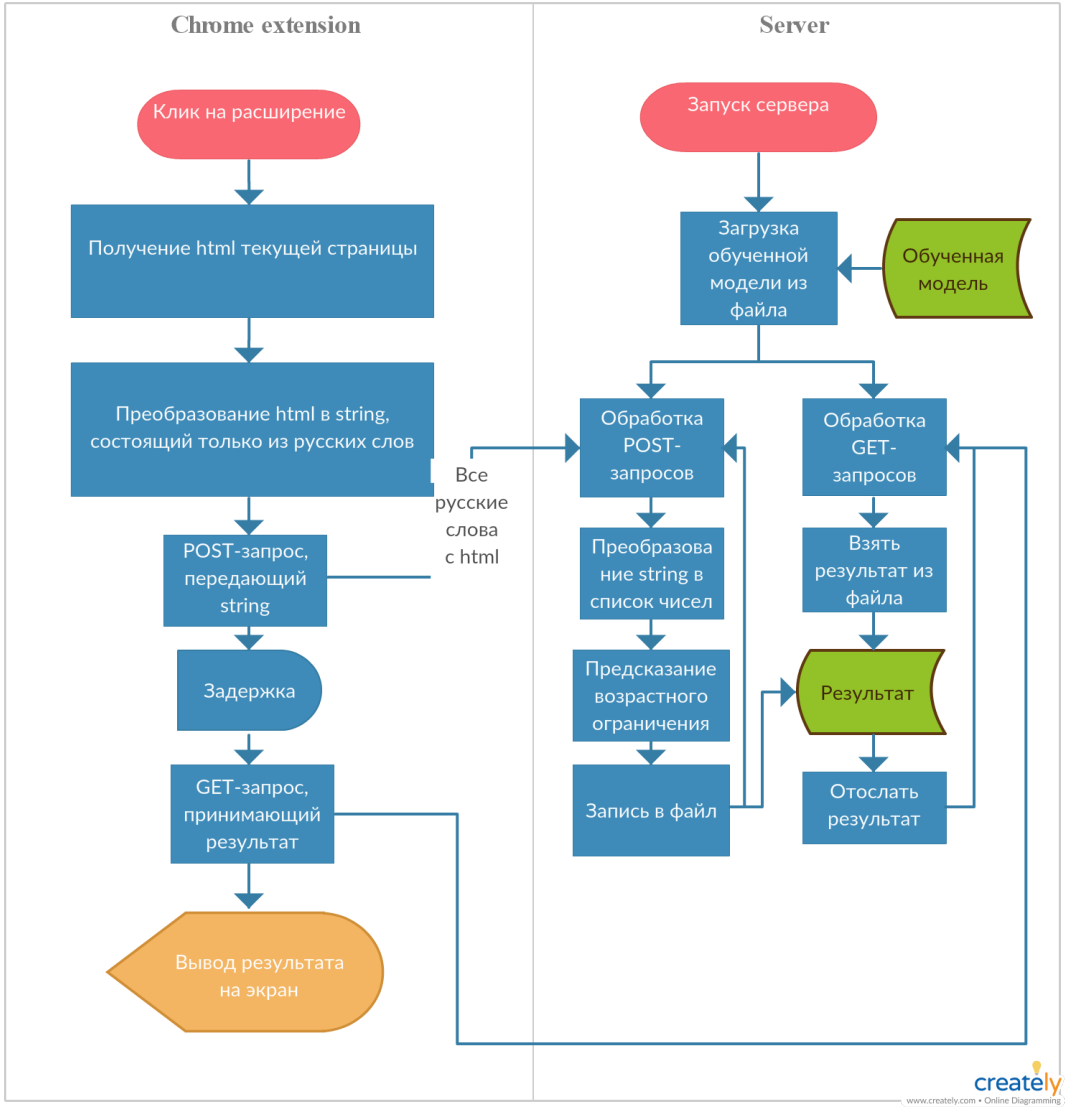
	F1 Score	Accuracy	Recall	Precision
Random model	0.58	0.51	0.58	0.58
Rule-based model	0.06	0.41	0.03	1.0
Classifier	0.90	0.88	0.93	0.87
Upgraded Classifier	**0.91**	0.90	0.92	0.91

Результаты обучения





Взаимодействие расширения и сервера



Итоги

- Федор
  - Сбор данных
  - Библиотека
- Александр
  - Классификация текста
  - Сервер
  - Расширение

Результаты

- Сделано расширение для Chrome – https://github.com/SmirnovAlexander/PoemClassifier
- Сделана библиотека на pypi – https://pypi.org/project/TalesParse/
- Собраны рассказы на kaggle – https://www.kaggle.com/idoldev/adult-and-child-russian-tales-dataset-with-label
- Написан сборщик рассказов – https://github.com/Feodoros/Scraping_Tales